Understanding new "exploratory" biomarker data: a first look at observed concentrations and associated detection limits

Joachim D. Pleil

National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

This editorial is the first of a series that each explains one practical aspect of statistics specifically tailored for biomarker data. Each editorial is focused on a very specific concept and gives the rationale, specific method, and a real-world example of a useful tool for data interpretation. The different topics are drawn from the author's experience in working with his own data and from teaching graduate students as to how and why certain statistical approaches can help answer real-world questions. The integration of simple and specific statistical tools in interpreting biomarkers information is an important contemporary issue for deducing relationships among environmental or clinical stressors, biomarkers, and ultimate health outcomes (Bean et al. 2015, Pleil and Sobus 2013). In addition, the ability to understand and compare disparate statistics is crucial to harmonize data across the literature (Pleil et al. 2014).

Here we focus on the first step in organizing new or exploratory biomarker data for which we do not have much prior knowledge. This occurs often in the biomarkers research community as instrumentation is becoming more sensitive and gives us access to lower concentrations of compounds in complex biological media; is generally referred to as "discovery" analysis (Pleil and Stiegel 2013). Although the discovery data sets are generally small and do not represent a broad spectrum of subjects/patients, there is a need to get some initial idea as to how the measured values relate to the practical analytical methods. In short, how well does the new method actually measure anything in real-world samples? As such, we experience two limitations:

- We do not know about the heterogeneity (distribution, range) of the biomarkers in real world samples
- We do not know, *a-priori*, how the instrument performance (detection limit) matches with the low-levels of biomarkers in real-world samples

The first issue of heterogeneity is general affected by meta-data grouping. It could be that males and females, old and young, etc. have different underlying distributions that cannot be well described with a few samples, which results in an erratic combined distribution. This behavior makes it difficult to properly assess the usual parametric statistics such as mean, standard deviation, confidence limits, etc. The second issue of "below detection limit" samples (left-censored data) makes direct comparisons within and between studies difficult; after all, without much prior knowledge, it is difficult to compare results when the underlying distributions have a lot of missing values.

These limitations call for an observational approach. Given a list of measurements data, we order them from lowest to highest in a spreadsheet (e.g. Excel), and then just look to see what measurement values are positioned as minimum, maximum, and the 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} (etc.) percentiles in the list. This requires no parametric estimates, just counting; in fact, excel provides a built in function that will do this for us: *=percentile(list, p)*, where "list" points to the column of numbers, and "p" is a number between 0 and 1 proportional to the percentile of interest. We do this for each of the analytes in the data set and construct a simple table.

For example, we had recently made measurements of a suite of 10 inflammatory cytokines using a new immunochemistry platform (MesoScale Discovery, MSD, Gaithersburg, MD) applied to samples from two biological matrices: human plasma, and exhaled breath condensate (EBC) (Stiegel et al. 2014). When we first embarked on this experiment, we only had estimates for instrument sensitivity based on synthetic samples, but we had no knowledge about actual cytokines expression in plasma or EBC in normal (healthy) adults. This was the exact situation where the discovery, or "observational" technique outlined above was appropriate. In Tables 1 and 2, below, we show the results from this experiment.

Table1 1: Descriptive Statistics (pg/mL) for Cytokines in Human Plasma

Cytokine	LLOQ[a]	%>LLOQ	Min	5%	25%	50%	75%	95%	Max
IL-8	0.056 (0.038)	100	0.980	1.15	1.50	1.97	2.67	3.86	41.7
TNF-α	0.064 (0.038)	100	1.06	1.29	1.44	1.73	2.49	3.73	8.77
IL-10	0.224 (0.060)	99.0	<LLOQ	0.486	0.746	1.26	1.59	2.31	3.23
IL-12 p70	0.197 (0.028)	95.2	<LLOQ	0.207	0.257	0.339	0.582	0.830	1.62
IL-5	0.055 (0.038)	94.3	<LLOQ	<LLOQ	0.115	0.208	0.373	0.881	20.4
IL-4	0.228 (0.117)	81.0	<LLOQ	<LLOQ	0.220	0.277	0.347	0.569	2.59
INF- γ	0.163 (0.104)	80.1	<LLOQ	<LLOQ	0.085	0.305	0.649	1.51	3.01
IL-13	1.34 (0.845)	78.1	<LLOQ	<LLOQ	0.220	1.62	2.04	2.75	4.38
IL-2	0.087 (0.068)	78.1	<LLOQ	<LLOQ	0.047	0.093	0.189	0.262	0.633
IL-1β	0.511 (0.717)	73.3	<LLOQ	<LLOQ	<LLOQ	0.160	1.56	3.38	40.3

[a] mean (standard deviation)

[*] LLOQ – estimate of lower level of quantitation

Table 2. Descriptive Statistics (pg/mL) for Cytokines in Human EBC

Cytokine	LLOQ[a]	%>LLOQ	Min	5%	25%	50%	75%	95%	Max
IL-10	0.369 (0.115)	93.5	<LLOQ	<LLOQ	0.581	0.737	1.78	2.38	3.15
IL-8	0.024 (0.011)	92.2	<LLOQ	<LLOQ	0.151	0.245	0.807	2.27	222
IL-4	0.143 (0.062)	90.9	<LLOQ	<LLOQ	0.323	0.411	2.92	3.82	4.16
IL-5	0.067 (0.071)	89.6	<LLOQ	<LLOQ	0.032	0.047	0.181	0.312	0.368
TNF-α	0.032 (0.023)	87.0	<LLOQ	<LLOQ	0.075	0.120	0.804	1.27	2.53
IFN-γ	0.035 (0.015)	81.8	<LLOQ	<LLOQ	0.058	0.099	0.582	1.22	1.62
IL-12 p70	0.172 (0.063)	72.7	<LLOQ	<LLOQ	<LLOQ	0.220	1.33	1.81	2.08
IL-1β	1.32 (0.365)	71.4	<LLOQ	<LLOQ	<LLOQ	2.90	5.27	13.2	94.8
IL-2	0.038 (0.027)	71.4	<LLOQ	<LLOQ	<LLOQ	0.061	0.659	1.06	2.31
IL-13	0.187 (0.126)	55.8	<LLOQ	<LLOQ	<LLOQ	0.244	2.86	3.64	4.26

[a] mean (standard deviation)
[*] LLOQ – estimate of lower level of quantitation

These simple tables tell us a great deal about cytokines in human plasma and EBC without invoking any detailed statistical calculations. We see immediately that the analytical technique is reasonably well suited for the task, and that we can get valid data for most subjects for most cytokines. The human plasma data are more complete as cytokines are at higher levels, but the breath data (EBC) show that it is also possible to use this non-invasive method to achieve a reasonable view of cytokine prevalence. We also see that there is room for improvement in analytical sensitivity if we want to capture data for all samples, especially below the 5th percentile in plasma, and the 25th percentile in EBC. Nonetheless, we can rank the 10-cytokines for prevalence and concentration in plasma and EBC, and use this information for future study designs.

The elegance of this approach is that the data come from an ordered list; there is no need for understanding the distribution, deciding on log-transformations, or any further calculations. Certainly, more detailed approaches that assess the underlying distributions (normal or lognormal) and apply appropriate parametric statistics are the ultimate goal, especially when we want to predict and compare mathematical relationships beyond the confines of the data set. However, as a first step for understanding new exploratory biomarkers data, a simple percentiles breakdown is an excellent observational tool.

Acknowledgements:

The author is thankful for expert advice and insights from colleagues, especially Dr. Andrew Lindstrom and Dr. Jon Sobus from US EPA, Dr. Michelle Angrish from ORISE/DOE and Dr. Matthew Stiegel from University of North Carolina, Chapel Hill. This work was reviewed by the U.S. EPA and approved for publication. Mention of trade names or commercial products does not constitute recommendation for use. The author declares he has no competing financial interests.

References:

Bean HD, Pleil JD, and Hill, JE, 2014. "Report: New analytical and statistical approaches for interpreting the relationships among environmental stressors and biomarkers", Biomarkers (in press).

Pleil JD, Sobus JR, Stiegel MA, Hu D, Oliver KD, Olenick C, Strynar MJ, Clark M, Madden MC, and Funk WE, 2014. Estimating common parameters of log-normally distributed environmental and biomonitoring data: harmonizing disparate statistics from publications, Journal of Toxicology and Environmental Health, Part B. Critical Reviews 17:341-368.

Stiegel MA, Pleil JD, Sobus JR, Morgan MK, and Madden MC, 2014. "Analysis of inflammatory cytokines in human blood, breath condensate, and urine using a multiplex immunoassay platform", Biomarkers, (in press).

Pleil JD and Stiegel MA, 2013. "The evolution of environmental exposure science: Using breath-borne biomarkers for "discovery" of the human exposome", Feature Article: Analytical Chemistry 85:9985-9990.

Pleil JD and Sobus JR, 2013. "Estimating lifetime risk from spot biomarker data and intra-class correlation coefficients (ICC)", Journal of Toxicology and Environmental Health A 76(12):747-766.